世界知的所有権機関
国際事務局

PCT

特許協力条約に基づいて公開された国際出願

(51) 国際特許分類6
G06F 17/30

A1

(11) 国際公開番号 WO99/14690

(43) 国際公開日 1999年3月25日(25.03.99)

(21) 国際出願番号 PCT/JP97/03280

(22) 国際出願日 1997年9月17日(17.09.97)

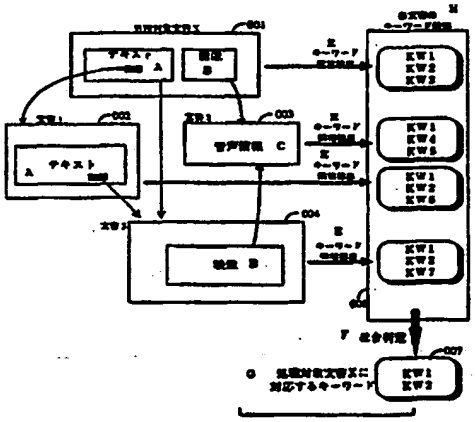
(71) 出願人（米国を除くすべての指定国について）
株式会社 日立製作所(HITACHI, LTD.)[JP/JP]
〒101 東京都千代田区神田駿河台四丁目6番地 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
間瀬久雄(MASE, Hisao)[JP/JP]
〒547 大阪府大阪市平野区長吉長原西3丁目10番33号
ACTY21 205号室 Osaka, (JP)
辻 洋(TSUJI, Hiroshi)[JP/JP]
〒664 兵庫県伊丹市車塚2丁目125番地 Hyogo, (JP)
(74) 代理人
弁理士 小川勝男(OGAWA, Katsuo)
〒100 東京都千代田区丸の内一丁目5番1号
株式会社 日立製作所内 Tokyo, (JP)

(81) 指定国 JP, US, 欧州特許 (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

添付公開書類
国際調査報告書

(54)Title: KEYWORD ADDING METHOD USING LINK INFORMATION

(54)発明の名称 リンク情報を用いたキーワード付与方法



(57) Abstract

Keywords proposed are extracted from an objective document and a document linked therewith, and they are integrated to qualify keywords of the objective document. The objective document is classified in a category by comparing these keywords with the processing keywords for classification. When a document is shown, the keywords concerning a document linked with the document or a document accessing frequency (an object corresponding thereto) is shown in an additionally arranged state.

(57)要約

対象文書およびそれにリンクした文書からキーワード候補を抽出し、それらを総合して対象文書のキーワードを認定する。また、このキーワードと分類知識中のキーワードとを照合することにより対象文書をカテゴリ分類する。また、文書を表示する際にその文書にリンクした文書に関するキーワードまたは文書アクセス頻度（に対応するオブジェクト）を付加配置して表示する。

明 細 書

リンク情報を用いたキーワード付与方法

技術分野

本発明は、ある文書及び関連する別の文書から、その内容を特徴付けるキーワードを自動抽出し、抽出したキーワードに基づいて文書を内容別に分類し、さらに、検索した文書の内容を表示する方法に関する。特に、ネットワーク上に散在する文書情報の中から適切なキーワードを抽出する方法に関する。

背景技術

文書情報を内容に応じてカテゴリに分類する方法としては、（１）文書からキーワードを抽出し、その出現傾向から適切なカテゴリを決定する方式が一般的である。

また、自分の所望の情報に効率良くアクセスするための方式としては、（２）ある文書に関連する文書をリンクさせておきリンクを適切に辿っていく方式や、（３）文書検索システムを利用する方式、すなわちキーワードや日付，作成者等に関連する検索条件をユーザに入力させ、その条件に合致する情報一覧を表示する方式、（４）文書情報ディレクトリを探索する方式、すなわちその内容に応じて各文書を予めジャンル分類しておき、ユーザにジャンル体系を探索させることにより文書の絞り込みを促進する方式がある。上記（４）のジャンル体系の探索方法としては、ジャンル体系の最上位から順に階層レベル別に表示していき、ユーザにトップダウンに辿らせる方法のほか、ジャンル体系一覧を全部表示して特定ジャンルを直接選択させる方法がある。

しかし、上記従来技術においては、以下の問題点が存在する。

（１）キーワードに基づく文書分類では、その文書にテキスト情報がある程度存在することが不可欠であるため、テキスト情報が全くあるいはあまり存在し

ない文書から適切なキーワードを抽出して、当該文書を内容別に分類することが不可能である。また、キーワードの抽出精度が分類精度に大きく影響するため、できるだけ多くの角度及び判定要素からキーワードを選定して抽出することができれば、それだけ高精度のキーワード抽出結果および分類結果が期待できる。

（２）ある文書情報に関連する文書がリンクされた文書群をリンクを辿って探索する場合、リンク先に記述された情報がユーザが所望する情報（あるいはユーザが所望する情報につながるパス上にある情報）であるかを判定するための手がかりは、リンク元のアンカー（別の文書を参照していることを表す語句）しかないため、ユーザが実際にリンクを辿って別の文書を見たときに必要な情報でなかったという場合が頻繁に起こっている。このような試行錯誤的な探索による検索効率の低下は、電話回線等によって情報にアクセスしているユーザに対して回線使用料等のコストが高くなるなどの問題をもたらす。

（３）ある情報に初めてアクセスする場合、ユーザはその情報がどんな内容や構成であるのか分かっていないので、キーワードを指定すること自体が困難である。また、検索条件の記述内容によっては、大量の検索結果が表示されることがあるため、検索条件を変えながら何度も検索をしなければならない恐れがあり、検索時間がかかる。

（４）ジャンルによる絞り込みについては、自分の要求する文書が当該ジャンル体系の中のどのジャンルに該当しているかを予め何らかの方法によって決定しなければならない。従って、選択したジャンルの中にユーザの要求する情報が含まれていない可能性がある。また、選択されなかったジャンルの中にユーザの要求する情報に関連する情報があった場合、要求した情報からさらにリンクを辿って、他のジャンルに属する関連情報に到達するための手がかりがないため、ここから後は試行錯誤的な探索を必要とし、上記（２）と同様の問題が生じる。

本発明の目的は、当該文書だけでなく、関連する文書を含む文書群から適切

なキーワードを抽出し、それに基づいて文書を高精度に分類する文書のキーワード付与方法及び装置を提供することにある。

発明の開示

本発明では、記憶装置に格納されたキーワード付与対象文書と当該キーワード付与対象文書に関連付けられている文書とからキーワードを抽出し、抽出したキーワードを当該キーワード付与対象文書に対応させて前記記憶装置に記憶させることにより、上記課題（１）を解決する。

また本発明では、分類対象文書と当該分類対象文書に関連付けられている文書とからキーワードを抽出し、当該キーワードと、記憶装置に記憶された「複数のカテゴリ群および各カテゴリに対応するキーワード群を記述した分類知識」中のキーワードとを照合することによりカテゴリ毎に類似度を算出し、類似度の高い一種類以上のカテゴリを当該分類対象文書に対応付けることにより、上記課題（１）を解決する。

さらに本発明では、分類対象文書と当該分類対象文書に関連付けられている文書とからキーワードを抽出し、当該キーワードと、記憶装置に格納された「ユーザ識別子および各ユーザ識別子に対応するキーワード群を記述した分類知識」中のキーワードとを照合することにより、当該分類対象文書が各ユーザの要求する文書であるか否かを判別し、要求する文書である場合、当該分類対象文書の内容あるいはアドレス情報を当該ユーザに通知することにより、上記課題（１）を解決する。

さらに本発明では、記憶装置に格納された文書と当該文書に関連付けられている文書とからそれぞれ一種類以上のキーワードを抽出して記憶装置に記憶しておき、前記文書を出力手段を介して表示する際に前記キーワードを、前記関連付けられている文書と対応するように配置して表示することにより、上記課題（２）、（３）、及び（４）を解決する。

さらに本発明では、記憶装置に格納された文書に関連付けられている文書が

アクセスされた回数を保持し、前記文書を出力手段を介して表示する際に当該表示対象文書とともに前記アクセス回数あるいはアクセス回数に対応するオブジェクトを文書毎に１対１に対応するように配置して表示することにより、上記課題（２）、（３）、及び（４）を解決する。

図面の簡単な説明

第１図は本実施例で述べるキーワード付与の概要を示す図であり、第２図は本実施例で述べるシステムの概要を示す図であり、第３図はリンク構造をなす文書群の一例を示す図であり、第４図は第３図の文書に関するＨＴＭＬ言語による記述例を示す図であり、第５図は単語重み付与ルール１７の構成の一例を示す図であり、第６図は文書解析処理部６の処理手順を示す図であり、第７図は単語テーブル１５の構成の一例を示す図であり、第８図はキーワード抽出ルール１８の構成の一例を示す図であり、第９図は文書および単語テーブル１５から認定されたキーワードの一例を示す図であり、第１０図はキーワード認定処理部７の処理手順を示す図であり、第１１図は分類知識ベース２０の定義の一例を示す図であり、第１２図は分類処理部８の処理手順を示す図であり、第１３図は文書情報データベース２２の構成の一例を示す図であり、第１４図はリンク情報挿入処理部１２の処理手順を示す図であり、第１５図はリンク情報挿入処理後のＨＴＭＬ言語による記述の一例を示す図であり、第１６図はリンク情報挿入処理後の文書表示結果の一例を示す図である。

発明を実施するための最良の形態

以下で、本発明の実施例を図面を用いながら詳細に説明する。

本実施例は、ネットワーク上に散在する文書群を収集して、各文書を特徴付けるキーワードを自動的に認定し、キーワードに基づいてこれらの文書を内容別に分類しておき、ユーザからの検索要求に合致する文書をユーザに表示するシステムである。検索の対象となる文書は、ある一つの文書から関連する別文

書にリンクをはることが可能なハイパーリンク構造を有していると仮定しており、本実施例では特に、WWW（World Wide Web）ブラウザによるアクセスが可能な、ＨＴＭＬ言語（Hypertext Markup Language）記述の文書とする。また、ＨＴＭＬでは、文字修飾やタイトル情報，リンクに関する情報を各種タグを用いて記述しているので、これらのタグの種類およびタグの範囲を解析することにより、さまざまな情報を抽出できる。さらに、ＨＴＭＬでは、画像情報，映像情報，音声情報を含めることが可能である。なお、本実施例で述べる内容は、この他にも各エンドユーザによる文書情報の分類整頓等にも適用可能である。

第１図は、本実施例の特徴を示す概要図であり、第２図以下で詳細に述べるシステムの理解を支援するための図である。

第１図において、処理対象（キーワード抽出、分類、表示等）文書００１から他の文書（００２，００３，００４）にリンクがはられており、互いに関連付けられている。本実施例では、まず後述する方法により、各文書からキーワード候補００６を抽出する。そして、それらのキーワード候補を総合的に評価することにより、処理対象文書００１に対するキーワード００７を認定する。つまり、処理対象文書に十分なテキスト情報が存在していなくても、その文書にリンクしている文書に含まれるキーワードであって、ある条件（後述）を満たすキーワードを処理対象文書のキーワードとすることにより、高精度のキーワード情報を出力できるのが本発明の効果の一つである。

また、第１図において、文書は、テキストの他に、音声，画像，映像情報を含むものもある。この場合、音声認識，画像処理，映像中の画像・音声認識処理を施すことにより、これらの情報からテキスト情報を抽出できるので、抽出後はテキストと同様に扱うことが可能である。

第２図は、本実施例で述べるシステムの概要を示す図である。

第２図のシステムは、文書情報が散在しているインターネット等の外部ネットワーク１，外部ネットワーク１から文書情報を収集して管理する文書管理サ

ーバ2，文書管理サーバ2に検索を要求して検索結果等をブラウザ２８に表示するクライアント3，文書管理サーバとクライアント群を連結するためのネットワーク４（ＬＡＮ（Local Area Network），電話回線等）から構成される。もちろん、文書管理サーバ2が収集管理する対象の文書は、ＬＡＮ４内のものを含んでも良い。

本実施例で述べるシステムは、以下の六つの機能を有する。

（１）文書情報を収集する。

（２）収集文書および当該文書に関連付けられた文書から、収集文書の各々を特徴付けるキーワードを認定する。

（３）当該キーワードに基づいて各文書を内容別に自動分類し、文書情報データベースに格納する。

（４）エンドユーザから要求される検索要求に従って文書ＤＢを検索し、検索結果をエンドユーザに返す。

（５）予めエンドユーザから自分の興味のある情報に関連するキーワードが指定されている場合、収集された文書の中にその興味に合致する情報があれば、その情報のアドレスをエンドユーザに報知する。

（６）エンドユーザからある文書へのアクセス・表示を要求された場合、取得した文書情報の中にその文書が参照している文書に対応するデータを添付してエンドユーザに表示する。

このうち、特に重要な機能は、（２）のキーワード認定であり、第２図における点線で囲まれた部分に相当する。

まず本システムでは、文書収集処理部５において、外部ネットワーク１に散在する文書情報を収集する。各文書にはアドレス情報が一意に決められている。WWWでは、ＵＲＬ（Uniformed Resource Locator）と呼ばれるアドレス情報が決められている。ＵＲＬには、その情報が格納されているサーバ名も含まれている。

一般に文書の収集は、あるページから始まって、そのページにリンクされて

いるページを辿っていくことによって行われる。文書収集のアルゴリズムについては既に公知であるのでここでは特に言及しない。なお、文書の収集は、自動的に収集するほかに、文書作成者自身が文書管理サーバ２の特定場所に文書を格納することにより収集する方法でも良い。文書収集処理部５で収集された文書は、文書データ１３に一時的に格納される。

次に、文書解析処理部６では、文書データ１３に収集蓄積された文書のテキスト部分から、当該文書を構成する単語を抽出する。文書がテキストでない（音声・画像・映像）場合、それぞれの情報からテキスト情報を認識するプログラムを適用することにより、テキスト情報を抽出する必要がある。音声認識，画像認識（特に文字認識）については、あるレベルの精度を持つシステムが既に実現されている。

文書解析処理６では、得られたテキスト文章を単語に分割するために、単語の見出しおよび品詞等の語彙情報を格納した単語辞書１６を参照する。単語分割アルゴリズムについては、情報処理学会第４４回全国大会講演論文集（３）１８１ページ等に示すように既に公知であるのでここでは言及しない。

また、文書解析処理部６では、各単語が当該文書においてどのくらい重要であるかを判定するために、単語重み付与ルール１７を参照することにより、各単語に重みを配分する。本実施例における単語重み付与ルール１７では、文章の記述に関する次の７種類のパラメータに関するルールを定義可能としている。

（１）文書のタイトル（ＨＴＭＬでは陽に現れない。タグ<TITLE>とタグ</TITLE>との間に記述される）

（２）文字の大きさ

（３）文字の色

（４）文字のスタイル（ゴシック，イタリック，アンダーライン等）

（５）出現頻度

（６）先頭からＮ文字以内に出現する語句

（７）他文書へのリンクを示すアンカー情報

各ルールには、当該ルールを満たす単語に加算すべき重みが定義されている。文書解析処理部６では、上記７種類のパラメータの少なくとも１種類を用いて、単語に重みを付与する。付与した結果は、単語テーブル１５に文書毎に格納される。本実施例では、単語テーブル１５には文書から抽出された名詞のみが格納され、残りは棄却される。

さらに、文書解析処理部６では、当該文書にリンクしているアンカーを認定することによって、当該文書にリンクしている他文書のＩＤ（ＵＲＬ）を認定する。ＨＴＭＬでは、アンカー情報は、「<A HREF="リンク先のアドレス">アンカー</A>」という方法で記述されるので、これを手がかりに"リンク先のアドレス（ＵＲＬ）およびアンカー文字列情報を容易に得ることができる。これらのリンク情報はリンク情報テーブル１４に対にして格納される。

次に、キーワード認定処理部７において、ある特定の文書のキーワードを、当該文書およびその文書にリンクされている文書群から抽出された単語（名詞）の中から総合的に決定する。すなわち、当該文書に出現する単語だけでなく、隣接する（関連する）文書に出現する単語の出現傾向をも踏まえたキーワード認定を行う。これにより、当該文書にテキスト情報がほとんど存在しない場合や、存在したとしても適切なキーワードの記述がない場合でも、関連する文書のキーワードを参照して総合的に判定することにより、当該文書に適切なキーワードを付与することが可能となる。

キーワードの認定は、キーワード抽出ルール１８を参照して行う。本実施例では、キーワードの認定を次の３種類のパラメータに従って行う。

（１）各文書から抽出された単語の持つ重みの値があるしきい値以上の単語。

（２）各文書から抽出された単語のうち、ある割合以上の文書（分類対象文書にリンクしている文書）に存在する単語。

（３）各文書から抽出された単語のうち、ある割合以下の文書（分類対象文書にリンクしている文書）に存在する単語。

しきい値および割合の値は、ルールの中で指定することができる。ルールと

して定義されたこれら３種類のパラメータの少なくとも１種類を満たす単語を当該文書のキーワードとして認定し、キーワードテーブル１９に格納する（もちろん、本システムは、文書からのキーワード抽出を目的とする場合でも適用可能である。その場合、キーワードテーブル１９に格納されたキーワードが最終出力となり、ここで処理は終了となる）。

次に、分類処理部８において、分類対象文書を予め定義されたカテゴリの少なくとも一つ以上に分類する。カテゴリは分類知識ベース２０の中で記述される。本実施例における知識ベース２０は、カテゴリの名称、各カテゴリに対応するキーワード，およびそのキーワードの重要度を示す重みの３要素が１組になって構成される。分類知識ベース２０は、カテゴリ毎にキーワードおよびその重みを定義することにより人手で作成しても良いし、各カテゴリに対応するサンプルテキストデータからキーワードおよびその重みを抽出することにより（半）自動的に生成しても良い。

分類処理部８では、キーワードテーブル１９に格納された分類対象文書のキーワードと、分類知識ベース２０に記述されたキーワードとを照合することにより、カテゴリ毎に類似度を算出する。これについては後で詳しく述べる。

各カテゴリ毎の類似度を算出した後、これらを類似度によりソートする。そして予め決められたしきい値以上の類似度を持つカテゴリを当該文書に付与する。しきい値を用いる代わりに、カテゴリの個数を決めておいても良いし、最大カテゴリ数Ｎを決めておき、しきい値以上の類似度を持つカテゴリのうちの上位Ｎ個のカテゴリを付与しても良い。分類結果として付与されたカテゴリは分類テーブル２１に格納される。

解析の終わった文書およびその属性情報（カテゴリなど）はすべて文書情報データベース２２に格納される。文書情報データベースには、文書ＩＤのほか、更新日（登録日）、分類処理部８で付与されたカテゴリ、キーワード認定処理部７で認定されたキーワードのほか、リンクしている文書ＩＤ、当該文書へのアクセス頻度、本文内容などが格納される。ここで、アクセス頻度情報につい

ては、後述するリンク情報挿入処理部１２からの更新要求により更新される。

本実施例における文書管理サーバ２は、複数のクライアント３からの検索に関連する要求を受理し、処理結果をクライアントに返すというクライアントーサーバシステム（ＣＳＳ）となっている。ＣＳＳの実現方式については既に公知であるので、ここでは説明しない。説明を簡単にするため、本実施例で述べるクライアント３からの要求内容は、次の３種類とする。実際には、他の要求があるであろう。

（１）ある条件を満たす文書を文書情報データベース２２から検索し、結果を取得する検索実行指示２６。

（２）あるアドレスに対応する文書情報を取得する文書アクセス指示２８。

（３）分類知識ベース２０の内容の定義・更新。

クライアント３から検索実行指示２６がなされると、検索に必要な検索条件が文書管理サーバ２の文書検索処理部１０にネットワーク４を経由して渡される。検索条件は、基本的には論理演算子（AND/OR）を用いることが一般的であり、キーワード，カテゴリを記述することによりユーザが作成する。文書検索処理部１０では、クライアント３から渡された検索条件を満たす文書情報を文書情報データベース２２から抽出する。論理式に基づいて文書を検索する方式については、例えば情報処理学会第４５回全国大会講演論文集（３）２３９ページ～２４４ページ等により公知なので、ここでは説明しない。

検索された文書情報のうち、文書ＩＤおよび更新日（登録日）のリストが検索結果２３に一時的に格納される。このリスト情報を検索要求をしてきたクライアント３にネットワーク４を経由して戻す。クライアント３では、返された文書ＩＤ情報をブラウザに表示する。

クライアント３からあるアドレスを持つ文書情報へのアクセス指示２７がなされるとユーザが入力したアドレス文字列を文書管理サーバ２のリンク情報挿入処理部１２にネットワーク４を経由して渡す。リンク情報挿入処理部１２では、当該アドレスに対応する情報を文書情報データベース２２，あるいは、キ

ャッシュディレクトリ，あるいは、内部ネットワーク４や外部ネットワーク１から取得する。ネットワーク上の文書情報を取得する方式については、ＵＲＬを指定してWWWの情報を取得することで既に実現されているので、ここで深く言及しない。

取得された文書情報について、当該文書情報が文書情報データベース２２の中に存在した場合、当該文書にリンクしている文書ＩＤおよびリンク先の文書ＩＤの持つキーワードを文書情報データベース２２から取得する。アクセス対象となっている文書において、どの文書とどこでリンクしているかに関するアンカー情報は、前述したように特定のタグを手がかりに認定できるので、リンク情報挿入処理部１２では、ある文書にリンクしているアンカーの直後に当該リンク先の文書に対応するキーワードを挿入する。キーワード情報の代わりに、文書情報データベースに格納されている当該文書へのアクセス頻度を挿入しても良い。

キーワードあるいはアクセス頻度といったリンク情報が挿入された文書はリンク情報付き文書２５に一時的に格納される。このデータは、ネットワーク４を経由してアクセス要求のあったクライアントに渡される。リンク情報挿入処理部１２では、アクセス要求がある度に、当該文書に対応するアクセス頻度を１ずつインクリメントするように文書情報データベース２２に要求する。

アクセス要求された文書情報が文書情報データベース２２にない場合は、上記リンク情報は表示しない。この文書ＩＤは、一時的に保持され、文書収集処理部５に渡して当該文書にリンクしている文書情報とともに収集し、キーワードを抽出して分類しておくことにより、次回以降は、リンク情報を添付することができる。

一方、本実施例を拡張することにより、あるユーザが自分の関心のある文書情報をユーザに報知することができる。すなわち、ユーザが自分の興味あるトピックに関連したキーワード（およびその重要度）を定義しておくと、本システムが新たに収集した文書あるいは既に収集した文書であって内容が更新され

ている文書について、上記方式により抽出されるキーワードと各ユーザの定義したキーワードとの間でマッチングを行って類似度を計算することにより、あるしきい値以上の類似度を有するキーワードを定義したユーザに対して当該文書のアドレス情報を電子メール等で送付することができる。この場合、分類処理部８では、カテゴリ別の類似度計算とユーザ別の類似度計算の両方を行う。ユーザによって定義されたキーワードおよび重み情報は、分類知識ベース２０にユーザＩＤに対応させて格納される。ユーザｉと当該文書との間の類似度の計算については、後述する。

文書が分類されて分類テーブル２１に一時的に格納されると、文書配信処理部１１では、分類テーブル２１に格納された文書ＩＤと報知すべきユーザＩＤの情報に基づいて、ユーザに報知すべき文書ＩＤリストを作成する。そして、ユーザに電子メール等により当該リストを送付する。送付の完了したものについて、対応する分類テーブル２１の内容は消去される。

以上に述べたように、本実施例のシステムによれば、収集した文書について、その文書自身およびその文書に関連した文書からキーワードを総合的に認定することができる。また、これらのキーワードを用いることにより、文書を分類したり、文書を表示する際に補足的なリンク情報（キーワード情報、アクセス頻度情報）をもユーザに提示したりすることができる。

以下では、第２図の処理の詳細について、具体例を用いて説明する。

第３図は、リンク構造をなす文書群の一例を示す図である。

第３図では、５種類の文書が互いにリンクにより関連付けされた構造をなしている。第３図で下線の文字（アンカー）が他文書へのリンクを表している。WWWブラウザでは、文書１の文字列「会社概要」をマウスで選択すると、リンクのはられた文書２を表示することができる。また、各文書を構成する文字について、その大きさやスタイル等を特定のタグを用いることにより変えることができる。

第４図は、第３図の文書に関するＨＴＭＬ言語による記述例を示す図である。

ＨＴＭＬ言語では、不等号（＜，＞）で囲まれるタグを使って、不等号で囲まれる文字を修飾したり、他文書へのリンク情報を記述する。各タグは、特定の機能にユニークに対応している。ＨＴＭＬ文書は、その書誌情報を表す部分と、本文を記述する部分とがある。前者は、タグＨＥＡＤで囲まれており、後者は、タグＢＯＤＹで囲まれている。書誌情報には文書のタイトル情報を記述することが可能である（この情報はブラウザには表示されない）。また、タグには、表示する文字の大きさを表すタグ（Ｈ１，Ｈ２，．．．）や、改行を示すタグ（Ｐ，ＢＲ）、他文書への参照を表すタグ（Ａ ＨＲＥＦ）などがある。それぞれのタグの機能は、そのタグにスラッシュ記号（／）を付けたタグが現れるまで有効であり、さらに、一部のタグについてはネストにすることも可能である。キーワードを抽出する際には、このタグ情報を参照しながらＨＴＭＬ文書を解析することになる。

第５図は、単語重み付与ルール１７の構成の一例を示す図である。

上述したように、原則的には、ある単語が重要であるか否かの判定はその単語に係るタグ情報を用いるので、本実施例では、単語重み付与ルール１７はタグ情報の有無に関するものとし、タグに対応させた重みを定義できるようにしている。

従って、第５図の第一レコードは、「ある単語がタグ「ＴＩＴＬＥ（文書タイトル情報）」の範囲内に出現するとき、その単語の重みに重み１０を加える」ことを表している。複数のタグ情報の範囲内にあるときは、それらすべてに対応する重みが加算される。なお、図５の最後のレコードの「Frequency」については、タグではなく、これは、文書内の単語の出現頻度に関するルールであり、予め指定された下限値以上の出現頻度の単語について重み３を加算するものである。下限値の設定の代わりに、文書内に出現する単語で出現率の高いものから上位Ｎ％というよう相対的な設定をしても良い。これらのしきい値情報は、この単語重み付与ルール１７に付加的に格納しておいても良いし、別の格納場所に格納しても良いし、処理プログラム内に埋め込んでも良い。

第６図は、文書解析処理部６の処理手順を示す図である。

文書解析処理部６では、ＨＴＭＬ文書の最後に到達するまで、以下の処理を行う（ステップ６００１）。まず、ＨＴＭＬ文書から文字列を１行読み取り（ステップ６００２）、その文字列をタグ情報と文章情報にわける（ステップ６００３）。次に、タグ情報について、タグ文字列の直前にスラッシュ（／）がついているか否かによって、タグが有効であるか否かを判定し（ステップ６００４）、有効である場合、そのタグ情報を保持する（ステップ６００５）。また、タグ情報がリンクを示すタグ「Ａ ＨＲＥＦ」であるかを判定し（ステップ６００６）、リンクを示すタグである場合、タグ「Ａ」に続くタグ「ＨＲＥＦ」という記述の直後に記述されている二重引用符で始まる文字列をリンク先文書のＩＤであると認定し、リンク情報テーブル１４にリンク元の文書ＩＤとともに格納する（ステップ６００７）。タグ以外の文字情報については、単語見出しおよび品詞・活用情報を格納した単語辞書１６を参照して単語分割を行う（ステップ６００８）。単語分割のアルゴリズムについては、最長一致法、最小コスト法などいくつかの方法が公知であり、これらの手法が適用できるので、ここでは説明しない。次に、分割された単語から名詞のみを抽出し、作業エリアに一時的に格納する（ステップ６００９）。次に、単語重み付与ルール１７を参照し、現時点での有効タグ情報の中に、ルール１７で指定されたタグが含まれている場合（ステップ６０１０）、ルール１７における当該タグに付与されている重みを、当該タグの有効範囲に存在する単語に付与する（ステップ６０１１）。

文書の最後まで解析が終わった後、作業エリアに格納されていた単語の出現頻度を単語別にカウントする（ステップ６０１２）。そして、各単語について（ステップ６０１３）、その出現頻度があるしきい値以上であるか否かを判定し（ステップ６０１４）、しきい値以上である場合、当該単語の重みに単語重み付与ルール１７に定義された重み（第５図ではFrequency の項目に該当する重み３）を加算する（ステップ６０１５）。ただし、ある特定の単語に対して

ある特定のタグに対応する重みを付与するのは一度限りとする。例えば、単語Aが文書中に２度イタリックで出現したとしても、単語Aに加算する、イタリックのタグに対応する重みは３（６ではない）とする。なお、タグ以外の情報に基づく重みの付与方法は、単語出現頻度以外に、文頭からＮ文字までに出現する単語に対してある重みを付与するというルールを用いても良いし、特定の文字列を伴う単語に対してある重みを付与するというルールを用いても良い。ここまでの処理で算出された単語（名詞）およびその重みを重みに基づいて降順に数値ソートし、その結果を単語テーブル１５に格納する（ステップ６０１６）。なお、ステップ６０１６で、単語テーブル１５に単語を格納する際に、格納対象の単語が予め指定しておいた単語群の中に含まれる場合、その単語を格納しないようにしてもよい。これにより、明らかにキーワードとなり得ない単語（日本語では例えば、「場合」「とき」など）を除去することが可能となる。

第７図は、単語テーブル１５の構成の一例を示す図であり、第３図の文書の各々から抽出した単語（名詞）およびその重みの一例（重みの高い上位の単語）を示す図である。文書１中の単語「鶴亀電機」は、第４図にも示すように、ＴＩＴＬＥに出現しており、また、文字が大きく（タグ「Ｈ１」）、さらにＢｏｌｄ体（タグ「Ｂ」）で記述されているので、第５図のルール１７を用いるとすると、その重みは、１０＋５＋７＝２２となる。同様に文書３中の単語「ＰＣ」については、ボールド体（タグ「Ｂ」）で出現し、文字が大きく（タグ「Ｈ２」）、文書４へのリンクを表すアンカー文字列を構成しているので、その重みは、５＋３＋８＝１６となる。

第８図は、キーワード抽出ルール１８の構成の一例を示す図である。

本実施例では、以下の条件を使って、各文書のキーワードを認定する。

（１）当該文書から抽出された単語について、その単語の重みがあるしきい値以上であるか。

（２）当該文書から抽出されたあるしきい値以上の重みを持つ単語について、

当該文書および当該文書からリンクがはられている文書のうちその単語の出現する文書数があるしきい値以上（以下）であるか。

（３）当該文書からリンクがはられている文書から抽出された単語について、その単語の重みがあるしきい値以上であるか。

（４）当該文書からリンクがはられている文書から抽出されたあるしきい値以上の単語について、当該文書および当該文書からリンクがはられている文書のうちその単語の出現する文書数があるしきい値以上（以下）であるか。

上記の条件を満たす単語を当該文書のキーワードとして認定する。第８図では、上記条件におけるしきい値情報を定義したものである。第８図では、上記条件（１）の重みのしきい値として１０と定義されていることを示している。また、上記条件（２）の重みのしきい値として５、文書数のしきい値として「６０％以上または１件以下」と定義されていることを示している。さらに、上記条件（３）の重みのしきい値として１５と定義されていることを示している。さらに、上記条件（４）の重みのしきい値として５、文書数のしきい値として「６０％以上または１件以下」と定義されていることを示している。

第９図は、第８図のキーワード抽出ルール１８に基づいて第７図の単語テーブル１５から認定された各文書のキーワードの一例を示す図である。文書１について説明すると、第８図の条件１よりまずキーワード「鶴亀電機」「ホームページ」が抽出される。次に条件２であるが、文書１の中には重みが１０以上でかつ出現する文書数の割合が６０％以上または１件以下の単語は存在しない（例えば「会社」はリンク先である文書２にも現れるがその出現文書数の割合は５０％（４件中２件）であり６０％に満たない）。次に条件３であるが、文書１にリンクしているのは文書２，３，４の３文書であり、この中で重みが１５以上である単語は、文書２の「会社」「概要」、文書３の「最新」「ニュース」、文書４の「製品」「情報」「インタネット」「対応」「ＰＣ」であるので、これらを文書１のキーワードとする。最後に条件４であるが、重みが５以上でかつ出現する文書数の割合が６０％以上または１件以下の単語は、「ハー

ドディスク」「プリンタ」であるので、これらを文書１のキーワードとする。結局、この例では、文書１のキーワードは、「鶴亀電機」「ホームページ」「会社」「概要」「最新」「ニュース」「製品」「情報」「インタネット」「対応」「ＰＣ」「ハードディスク」「プリンタ」の１３種類であると認定する。ここで、「ＰＣ」「インタネット」などは、文書１には現れないキーワードであることに注意されたい。なお前述したが上記キーワードのうち、「ホームページ」「概要」「最新」「対応」「情報」などは文書の内容を特徴付けるキーワードとしてはあまり適切でないと思われるので、このような単語リストを予め用意しておき、除去することは可能である。

第１０図は、キーワード認定処理部７の処理手順を示す図である。

まず、リンク情報テーブル１４からキーワード認定対象文書にリンクしている文書ＩＤを取得し、作業エリアに格納する（ステップ７００１）。次に、ステップ７００１で取得した文書に対応する単語およびその重みを単語テーブル１５からすべて取得し、作業エリアに格納する（ステップ７００２）。次に、作業エリアに格納された各単語について（ステップ７００３）、当該単語を含む文書が何件あるかをカウントし、作業エリアに格納された文書数に占める割合を算出し保持する（ステップ７００４）。次に、キーワード抽出ルール１８を参照して、条件 COND1 が定義されているか否かを判定し（ステップ７００５）、定義されている場合、作業エリアに格納されたキーワード認定対象文書の各単語について（ステップ７００６）、その重みが条件 COND1 に記述されたしきい値以上であるか否かを判定し（ステップ７００７）、しきい値以上である場合、当該単語を当該キーワード認定対象文書のキーワードとして、その文書ＩＤおよびその単語の重みとともにキーワードテーブル１９に格納する（ステップ７００８）。次に同様にして条件 COND2 が定義されているか否かを判定し（ステップ７００９）、定義されている場合、作業エリアに格納されたキーワード認定対象文書の各単語について（ステップ７０１０）、ステップ７００４で算出した値を参照しながら、当該単語の出現する文書数およびその

全体に占める割合がCOND2に記述された範囲を満たすか否かを判定し（ステップ７０１１）、満たす場合、当該単語を当該キーワード認定対象文書のキーワードとして、その文書ＩＤおよびその単語の重みとともにキーワードテーブル１９に格納する（ステップ７０１２）。次に、ステップ７００５からステップ７０１２と同様の処理を、当該キーワード認定対象文書にリンクしている文書の単語について行う（ステップ７０１３～ステップ７０２０、ただし、上記COND1の代わりにCOND3が適用され、上記COND2の代わりにCOND4が適用される）。本実施例では、４種類のキーワード抽出ルールを用いているが、これらはルールの一例であり、同様にして別のルールを定義することが可能である。

第１１図は、分類知識ベース２０の定義の一例を示す図である。

分類知識ベース２０は、使用目的の異なる２種類のテーブルから構成される。すなわち、文書をカテゴリに分類するためのカテゴリ分類テーブルと、文書をその内容に興味を持つユーザに対応付けるためのユーザ分類テーブルである。第１１図に示すように、前者はカテゴリ名，キーワード，重みの３種類から構成され、後者はユーザＩＤ，キーワード，重みの３種類から構成される。両者はカテゴリ名がユーザＩＤとなっているだけで、その他については同一の構成をしている。

カテゴリ分類テーブルは、本システムの管理者が手作業で定義することもできるし、当該カテゴリに該当するテキストを収集しておき、それらのテキストから本実施例で述べたような方式等によりキーワードを自動抽出することにより、（半）自動的に定義することも可能である。どちらの方法によって作成されてもかまわないが、とにかくカテゴリ分類テーブルが定義されていることは不可欠である。

また、ユーザ分類テーブルは、各ユーザがエディタ等により定義するものである。ただし、この場合、分類処理部８でキーワードの照合ができるように、ユーザが指定した単語は、単語辞書１６を参照して単語分割しておく必要があ

る。この際、ユーザが指定した単語が単語辞書１６に存在しない場合、その単語は適切に分割されることになる。

分類知識ベース２０に記述される重みは、数値が高いほど重要であるとする。この数値は、相対的数値（例えば、０から１の間）で記述しても良いし、絶対的数値（例えば、３０とか２００とか）で記述しても良い。第１１図では、前者を採用している。

第１２図は、分類処理部８の処理手順を示す図である。

カテゴリ（またはユーザ）毎の類似度の値を格納する配列要素を０に初期化した（ステップ８００１）後、キーワードテーブル19に格納された分類対象文書のキーワードについて（ステップ８００２）、分類知識ベース２０を参照して当該キーワードを持つカテゴリ（またはユーザＩＤ）に対して、次の値を計算し、当該カテゴリの類似度に追加する（ステップ８００３）。

$W_j \times (w_{ij} / \Sigma w_j)$

ここで、W_jは、当該キーワード（ｊ）の持つ重みの値をさす。w_{ij}は、知識ベース２０においてあるカテゴリｉに対応する当該キーワード（ｊ）の重みをさす。Σw_jは、当該キーワードについてのすべてのカテゴリの重みの合計をさす。

上式によれば、類似度計算は次の二つの性質をもつことになる。

（１）分類対象文書のキーワードの重みW_jが大きいほど類似度は大きくなる。

（２）あるカテゴリｉに対応するキーワードの重みの相対的割合（$w_{ij} / \Sigma w_j$）が大きいほど類似度は大きくなる。

なお、上式に代わる類似度計算方法として、当該キーワードの重みと、対応するカテゴリとの積を用いてもよい。また、これらの値に対して、単項演算子（log、√、べき乗、階乗など）を施したものを類似度としても良い。

次に、ここまでで算出された各カテゴリ（ユーザＩＤ）毎の類似度について、あるしきい値よりも大きな類似度を持つところのカテゴリを、分類テーブルに当該文書ＩＤとともに格納する（ステップ８００４）。

第１３図は、文書情報データベース２２の構成の一例を示す図である。

文書情報格納処理部９では、ある文書に関する各種データを文書情報データベース２２に格納する。文書情報データベース２２は、ユーザからの要求があったときに、文書検索処理部１０を介してそのデータ内容にアクセスされる。本実施例の文書情報データベース２２は、文書ＩＤ，更新日，カテゴリ，キーワード，アクセス頻度（初期値は０），リンク先文書ＩＤリスト，本文から構成される。

第１４図は、リンク情報挿入処理部１２の処理手順を示す図である。

ユーザからアクセス要求のあった文書ＩＤを受け取ると、まず、その文書情報を収集する（ステップ１２００１）。文書情報データベース２２から抽出しても良いが、文書の内容が更新されていることもあるので、ここでは、文書情報が格納されたサーバからネットワーク経由で文書情報を取得する。次に、当該文書ＩＤに対応するデータが文書情報データベース２２に存在するか否かを判定し（ステップ１２００２）、存在するならば、当該文書ＩＤにリンクする文書ＩＤおよびそのキーワード、当該文書ＩＤへのアクセス頻度情報を取得する（ステップ１２００３）。次に、当該文書のＨＴＭＬファイルを探索し、他文書へのリンクを示すアンカーを、リンクを示すタグを手がかりに順次見つけ、当該アンカーの直後にキーワード群あるいはアクセス頻度情報を挿入する（ステップ１２００４）。そして、文書情報データベース２２の当該文書のアクセス頻度の数値に１を加える（ステップ１２００５）。対応するデータが存在しない場合は、キーワード情報やアクセス頻度情報は挿入せず、そのままクライアントに送られる（ステップ１２００６）。もちろん、文書解析処理部６に渡してキーワード抽出を行っても良いが、解析にかかる処理時間の大小により、アクセス時間が増加すると考えられるので、本実施例では、とりあえずキーワード情報やアクセス頻度情報を挿入しないこととする。ただし、これら文書情報データベース２２に格納されていない文書ＩＤ情報を蓄積しておき、後にバッチで処理してデータベース２２に登録することはできる。

キーワード情報あるいはアクセス頻度情報の挿入された文書情報は、アクセス要求のあったクライアントに送られ、ブラウザ上に表示される。

第１５図は、第３図の文書に関するリンク情報挿入処理後のＨＴＭＬ言語による記述の一例を示す図である。アンカー文字列「最新ニュース」の直後に、このアンカーによりリンクされている文書３に関するキーワードおよび文書３へのアクセス頻度情報が挿入されている。なお、これらの情報を挿入するか否かは、ユーザが指定することも可能である。

第１６図は、第３図の文書に関するリンク情報挿入処理後の文書表示結果の一例を示す図である。各アンカーには、キーワードおよびアクセス頻度を示す情報が付加された形で表示されている。これにより、ユーザは、次にどのアンカーを辿れば所望の情報にたどり着けるのかを知ることができる。また、キーワードだけではどのリンクを辿ったらよいのか分からない場合には、アクセス頻度情報を参照することにより、他のユーザがより高頻度でアクセスしている文書からとりあえずアクセスしてみることができる。なお、キーワードの表示については、アンカーを構成する単語と、アンカーによりリンクされている文書のキーワードとの間に重複が見られることがある。この場合、リンク情報挿入処理部１２において、重複キーワードの除去をしても良い。

産業上の利用可能性

本発明によれば、ある文書からキーワードを抽出したり、ある文書を分類する際に、その文書内の情報だけでなく、その文書に関連付けられた文書情報から抽出したキーワードをも用いるので、当該文書に適切なキーワードが存在しない場合でも的確にキーワードを認定でき、高精度に文書分類することができる。また、本発明によれば、文書内容を表示する際に、当該文書にリンクしている文書に関するキーワード情報あるいはその文書のアクセス頻度情報を付加して表示するので、的確にリンクを辿っていくことができる。

これらにより、ユーザの所望する文書へ効率良くアクセスすることができ、検索時間および検索費用などのコストを低減できる。

請 求 の 範 囲

1．記憶装置に格納されたキーワード付与対象文書と当該キーワード付与対象文書に関連付けられている文書とからキーワードを抽出し、抽出したキーワードを当該キーワード付与対象文書に対応させて前記記憶装置に記憶させることを特徴とするリンク情報を用いたキーワード付与方法。

2．キーワード付与対象文書と当該キーワード付与対象文書に関連付けられている文書とから抽出したキーワードを当該キーワード付与対象文書に対応させて記録させたことを特徴とするコンピュータ読み取り可能な記録媒体。

3．前記キーワード付与対象文書は、音声データ，映像データ，画像データ，およびテキストデータの少なくとも一つを含むことを特徴とする請求の範囲第1項記載のリンク情報を用いたキーワード付与方法。

4．前記キーワード付与対象文書および当該キーワード付与対象文書に関連付けられている文書の各々から、（1）当該文書のタイトルを構成する語句、（2）他の文字に比べて文字の大きい語句、（3）他の語句と表示色の異なる語句、（4）他の語句と文字のスタイルが異なる語句、（5）出現頻度の高い語句、（6）特定の条件を満たす位置に出現する語句、（7）他の文書へのリンクを示す要素（アンカー）を構成する語句、のうちの少なくとも一つに関する語句抽出条件を満たす語句を当該文書に対応するキーワード候補とすることを特徴とする請求の範囲第1項記載のリンク情報を用いたキーワード付与方法。

5．前記語句抽出条件の各々に予め重みを定義しておき、ある抽出条件を満たす語句に当該抽出条件に対応する重みを加算し、予め指定されたしきい値以上の重みを持つ語句を当該文書に対応するキーワード候補とすることを特徴とする請求の範囲第4項記載のリンク情報を用いたキーワード付与方法。

6．前記文書の各々から抽出された前記キーワード候補から当該キーワード付与対象文書に対応するキーワードを認定する際に、（1）あるしきい値以上の重みを持つ語句、（2）抽出された語句のうち予め指定された割合以上の文書

に存在する語句、（３）抽出された語句のうち予め指定された割合以下の文書にのみ存在する語句、の少なくとも一つのキーワード認定条件を満たす語句をキーワードと認定し、当該キーワード付与対象文書に対応付けることを特徴とする請求の範囲第１項記載のリンク情報を用いたキーワード付与方法。

７．分類対象文書と当該分類対象文書に関連付けられている文書とから抽出したキーワードと、記憶装置に記憶されたカテゴリごとにキーワードを分類した分類知識中のキーワードとを照合することによりカテゴリ毎に類似度を算出し、類似度の高い一種類以上のカテゴリを当該分類対象文書に対応付けることを特徴とする文書分類方法。

８．分類対象文書と当該分類対象文書に関連付けられている文書とから抽出したキーワードと、記憶装置に記憶されたカテゴリごとにキーワードを分類した分類知識中のキーワードとを照合することによりカテゴリ毎に類似度を算出し、類似度の高い一種類以上のカテゴリを当該分類対象文書に対応させて前記記憶装置に記録させたことを特徴とするコンピュータ読み取り可能な記録媒体。

９．前記文書は、音声データ，映像データ，画像データ，およびテキストデータの少なくとも一つを含むことを特徴とする請求の範囲第７項記載の文書分類方法。

１０．前記分類対象文書および当該分類対象文書に関連付けられている文書の各々から、（１）当該文書のタイトルを構成する語句、（２）他の文字に比べて文字の大きい語句、（３）他の語句と表示色の異なる語句、（４）他の語句と文字のスタイルが異なる語句、（５）出現頻度の高い語句、（６）特定の条件を満たす位置に出現する語句、（７）他の文書へのリンクを示す要素（アンカー）を構成する語句、のうちの少なくとも一つに関する語句抽出条件を満たす語句を当該文書に対応するキーワード候補とすることを特徴とする請求の範囲第７項記載の文書分類方法。

１１．前記語句抽出条件の各々に予め重みを定義しておき、ある抽出条件を満たす語句に当該抽出条件に対応する重みを加算し、予め指定されたしきい値以

上の重みを持つ語句を当該文書に対応するキーワード候補とすることを特徴とする請求の範囲第１０項記載の文書分類方法。

１２．前記文書の各々から抽出された前記キーワード候補から当該分類対象文書に対応するキーワードを認定する際に、（１）あるしきい値以上の重みを持つ語句、（２）抽出された語句のうち予め指定された割合以上の文書に存在する語句、（３）抽出された語句のうち予め指定された割合以下の文書にのみ存在する語句、の少なくとも一つのキーワード認定条件を満たす語句をキーワードと認定し、当該分類対象文書に対応付けることを特徴とする請求の範囲第７項記載の文書分類方法。

１３．分類対象文書と当該分類対象文書に関連付けられている文書とから抽出したキーワードと、記憶装置に格納されたユーザ識別子ごとにキーワードを分類した分類知識中のキーワードとを照合することにより、当該分類対象文書が各ユーザの要求する文書であるか否かを判別し、要求する文書である場合、当該分類対象文書の内容あるいはアドレス情報を当該ユーザに通知することを特徴とする文書分類方法。

１４．記憶装置に格納された文書と当該文書に関連付けられている文書とからそれぞれ一種類以上のキーワードを抽出して記憶装置に記憶しておき、前記文書を出力手段を介して表示する際に前記キーワードを、前記関連付けられている文書と対応するように配置して表示することを特徴とする文書表示方法。

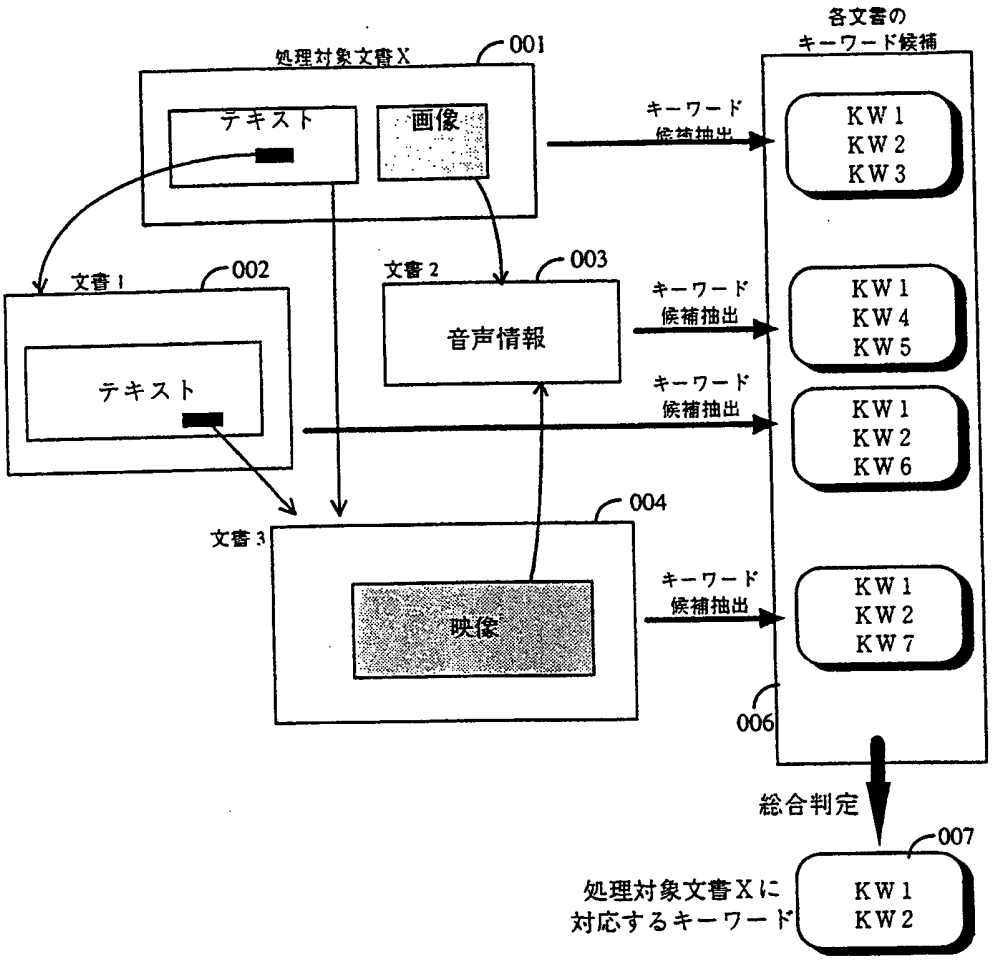
１５．前記関連付けられている文書の各々から前記キーワードを抽出する際に、（１）当該文書のタイトルを構成する語句、（２）他の文字に比べて文字の大きい語句、（３）他の語句と表示色の異なる語句、（４）他の語句と文字のスタイルが異なる語句、（５）出現頻度の高い語句、（６）特定の条件を満たす位置に出現する語句、（７）他の文書へのリンクを示す要素（アンカー）を構成する語句、のうちの少なくとも一つの語句抽出条件を満たす語句をキーワードとすることを特徴とする請求の範囲第１４項記載の文書表示方法。

１６．前記語句抽出条件の各々に対応して予め重みを定義しておき、ある抽出

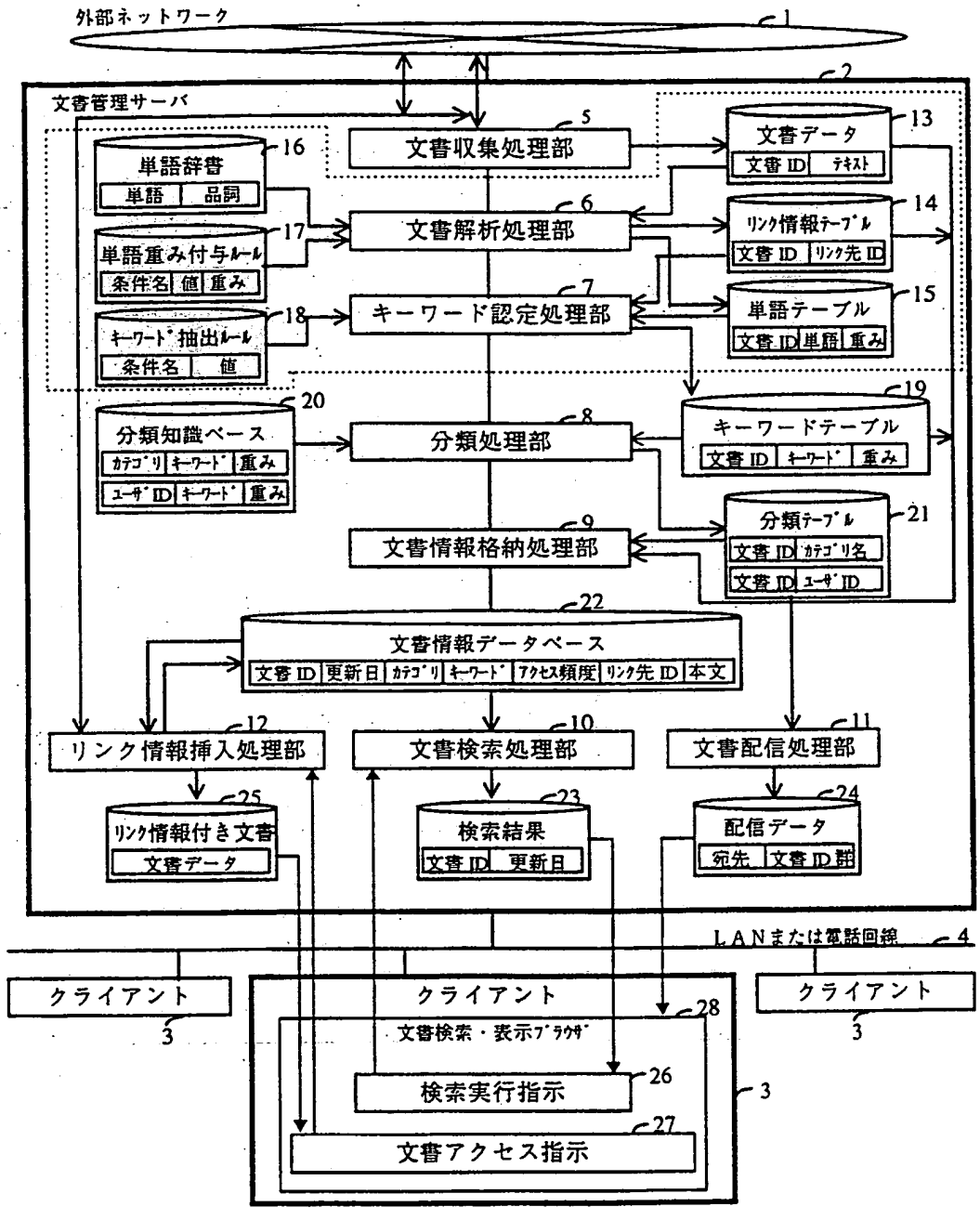
条件を満たす語句に当該抽出条件に対応する重みを加算し、予め指定されたしきい値以上の重みを持つ語句を当該文書のキーワードとすることを特徴とする請求の範囲第１４項記載の文書表示方法。

１７．記憶装置に格納された文書に関連付けられている文書がアクセスされた回数を保持し、前記文書を出力手段を介して表示する際に当該表示対象文書とともに前記アクセス回数あるいはアクセス回数に対応するオブジェクトを文書毎に１対１に対応するように配置して表示することを特徴とする文書表示方法。

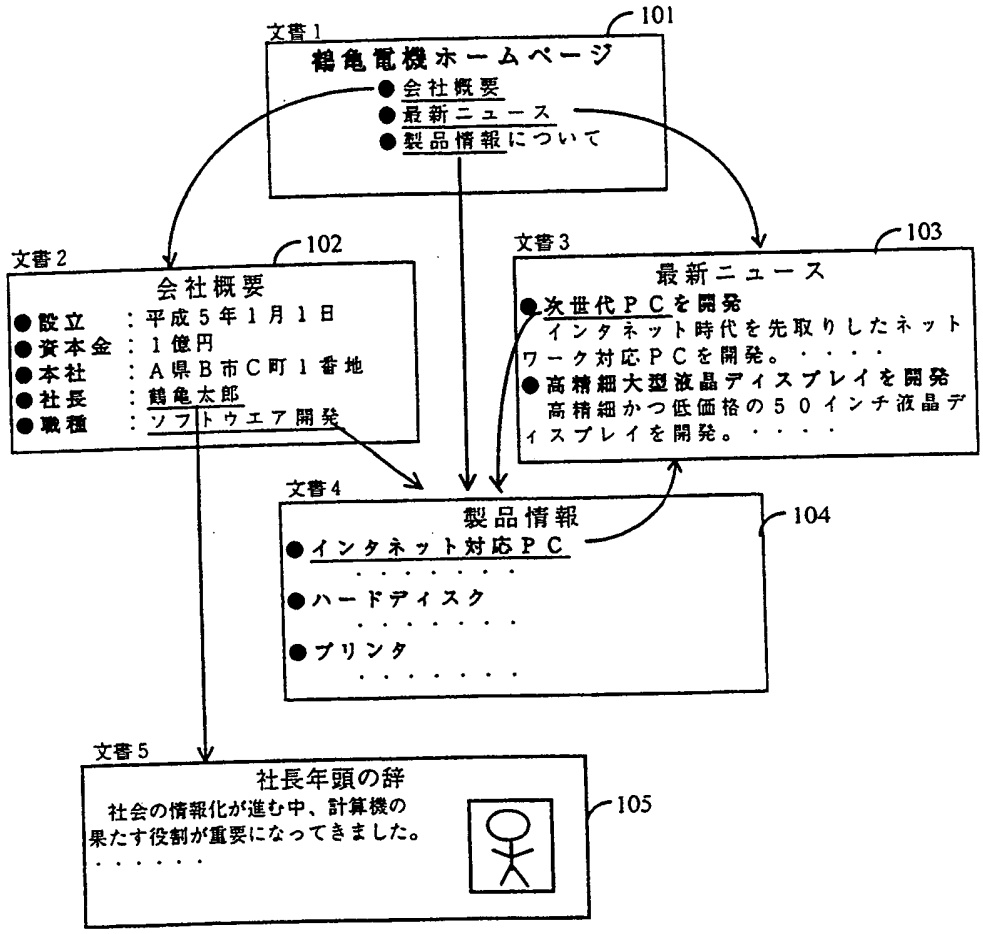
第 1 図


各文書の
キーワード候補
処理対象文書X
001
テキスト
画像
キーワード
候補抽出
KW1
KW2
KW3
文書1
002
テキスト
文書2
003
音声情報
キーワード
候補抽出
KW1
KW4
KW5
キーワード
候補抽出
KW1
KW2
KW6
文書3
004
映像
キーワード
候補抽出
KW1
KW2
KW7
006
総合判定
007
処理対象文書Xに
対応するキーワード
KW1
KW2

第 2 図


外部ネットワーク
1
2
文書管理サーバ
5
文書収集処理部
6
文書解析処理部
7
キーワード認定処理部
8
分類処理部
9
文書情報格納処理部
10
文書検索処理部
11
文書配信処理部
12
リンク情報挿入処理部
13
文書データ
文書ID テキスト
14
リンク情報テーブル
文書ID リンク先ID
15
単語テーブル
文書ID 単語 重み
16
単語辞書
単語 品詞
17
単語重み付与ルール
条件名 値 重み
18
キーワード抽出ルール
条件名 値
19
キーワードテーブル
文書ID キーワード 重み
20
分類知識ベース
カテゴリ キーワード 重み
ユーザID キーワード 重み
21
分類テーブル
文書ID カテゴリ名
文書ID ユーザID
22
文書情報データベース
文書ID 更新日 カテゴリ キーワード アクセス頻度 リンク先ID 本文
23
検索結果
文書ID 更新日
24
配信データ
宛先 文書ID群
25
リンク情報付き文書
文書データ
LANまたは電話回線
4
クライアント
3
クライアント
3
クライアント
28
文書検索・表示ブラウザ
26
検索実行指示
27
文書アクセス指示
3

第３図


文書1
101
鶴亀電機ホームページ
●会社概要
●最新ニュース
●製品情報について
文書2
102
会社概要
●設立 ：平成５年１月１日
●資本金 ：１億円
●本社 ：Ａ県Ｂ市Ｃ町１番地
●社長 ：鶴亀太郎
●職種 ：ソフトウエア開発
文書3
103
最新ニュース
●次世代ＰＣを開発
インタネット時代を先取りしたネットワーク対応ＰＣを開発。・・・・
●高精細大型液晶ディスプレイを開発
高精細かつ低価格の５０インチ液晶ディスプレイを開発。・・・・
文書4
製品情報
104
●インタネット対応ＰＣ
・・・・・・・
●ハードディスク
・・・・・・・
●プリンタ
・・・・・・・
文書5
社長年頭の辞
社会の情報化が進む中、計算機の果たす役割が重要になってきました。
・・・・・・
105

第４図

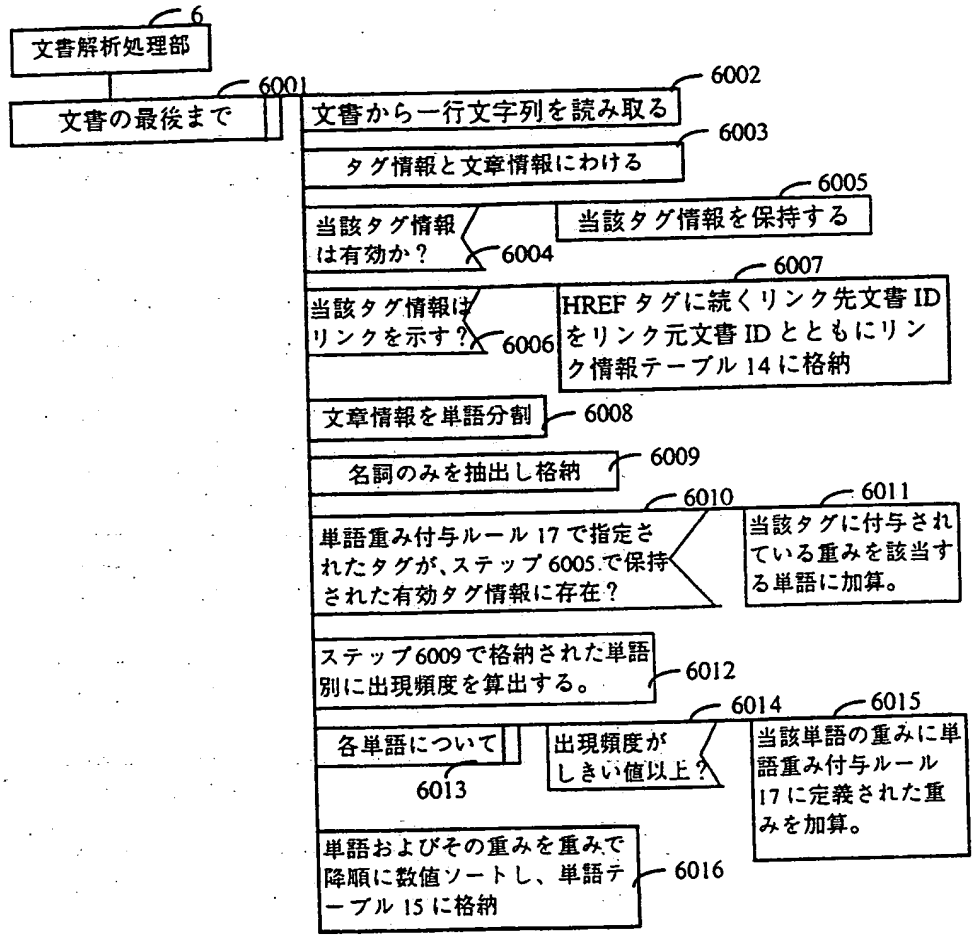
131

```
<HTML>
<HEAD>
<TITLE> 鶴亀電機のホームページ</TITLE>
</HEAD>
<BODY>
<H1><B>鶴亀電機ホームページ</B><P></H1>
<A HREF="文書２">●会社概要</A><BR>
<A HREF="文書３">●最新ニュース</A><BR>
<A HREF="文書４">●製品情報について</A><BR>
</BODY>
</HTML>
```

第５図

項目（タグ） (171)	重み (172)
TITLE （文書タイトル）	１０
I （イタリック）	3
B （ボールド）	5
U （アンダーライン）	3
EM （強調文字）	5
STRONG （強調文字）	5
H1（見出し文字の大きさ）	7
H2（見出し文字の大きさ）	3
H3（見出し文字の大きさ）	1
A HREF （ｱﾝｶｰ（他文書参照））	8
Frequency （単語出現頻度）	3

第 6 図


6
文書解析処理部
6001
文書の最後まで
6002
文書から一行文字列を読み取る
6003
タグ情報と文章情報にわける
6005
当該タグ情報を保持する
当該タグ情報は有効か？
6004
6007
HREF タグに続くリンク先文書 ID をリンク元文書 ID とともにリンク情報テーブル 14 に格納
当該タグ情報はリンクを示す？
6006
文章情報を単語分割
6008
名詞のみを抽出し格納
6009
6010
単語重み付与ルール 17 で指定されたタグが、ステップ 6005 で保持された有効タグ情報に存在？
6011
当該タグに付与されている重みを該当する単語に加算。
ステップ 6009 で格納された単語別に出現頻度を算出する。
6012
6014
6015
各単語について
6013
出現頻度がしきい値以上？
当該単語の重みに単語重み付与ルール 17 に定義された重みを加算。
単語およびその重みを重みで降順に数値ソートし、単語テーブル 15 に格納
6016

第７図

151 文書ＩＤ	152 単語	153 重み
文書１	鶴亀電機	２２
文書１	ホームページ	２２
文書１	会社	8
文書１	概要	8
文書１	最新	8
文書１	ニュース	8
文書１	製品	8
文書１	情報	8
・・・	・・・	・・・
文書２	会社	１７
文書２	概要	１７
文書２	鶴亀太郎	8
文書２	ソフトウエア	8
文書２	開発	8
・・・	・・・	・・・
文書３	最新	１７
文書３	ニュース	１７
文書３	次世代	１３
文書３	ＰＣ	１３
文書３	開発	１３
・・・	・・・	・・・
文書４	製品	１７
文書４	情報	１７
文書４	インタネット	１６
文書４	対応	１６
文書４	ＰＣ	１６
文書４	ハードディスク	8
文書４	プリンタ	8
・・・	・・・	・・・
文章５	社長	7
文章５	年頭の辞	7
・・・	・・・	・・・

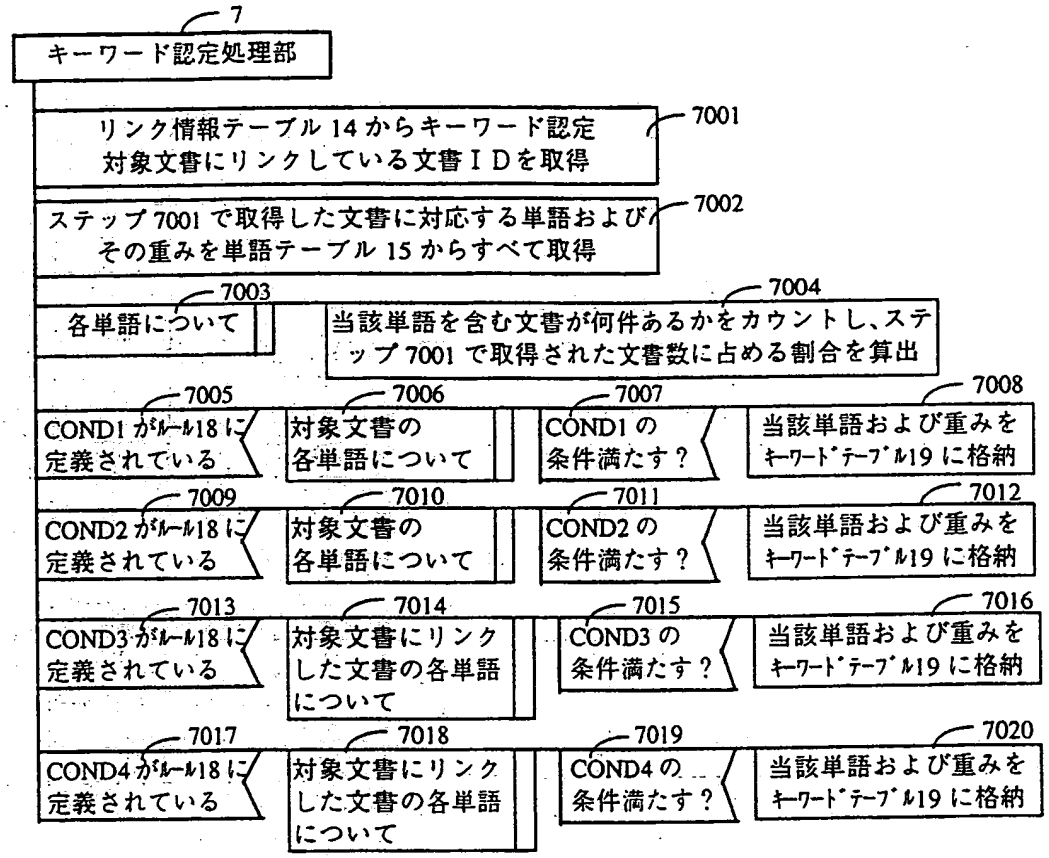
第 8 図

181 条件ID	182 条件項目	183 値
COND1	当該文書の単語の重みしきい値	15
COND2	(1)当該文書の単語の重みしきい値 (2)当該文書または当該文書にリンクした文書において、当該単語の出現する文書数	5 60%以上 または1件
COND3	当該文書にリンクした文書の単語の重みしきい値	15
COND4	(1)当該文書にリンクした文書の単語の重みしきい値 (2)当該文書または当該文書にリンクした文書において、当該単語の出現する文書数	5 60%以上 または1件

第９図

191 文書ＩＤ	192 キーワード	193 重み
文書１	鶴亀電機	２２
文書１	ホームページ	２２
文書１	会社	１７
文書１	概要	１７
文書１	最新	１７
文書１	ニュース	１７
文書１	製品	１７
文書１	情報	１７
文書１	インタネット	１６
文書１	対応	１６
文書１	ＰＣ	１６
文書１	ハードディスク	８
文書１	プリンタ	８
文書２	・・・	・・・
・・・	・・・	・・・

第１０図


7
キーワード認定処理部
リンク情報テーブル14からキーワード認定対象文書にリンクしている文書ＩＤを取得
7001
ステップ7001で取得した文書に対応する単語およびその重みを単語テーブル15からすべて取得
7002
各単語について
7003
当該単語を含む文書が何件あるかをカウントし、ステップ7001で取得された文書数に占める割合を算出
7004
COND1がルール18に定義されている
7005
対象文書の各単語について
7006
COND1の条件満たす？
7007
当該単語および重みをキーワードテーブル19に格納
7008
COND2がルール18に定義されている
7009
対象文書の各単語について
7010
COND2の条件満たす？
7011
当該単語および重みをキーワードテーブル19に格納
7012
COND3がルール18に定義されている
7013
対象文書にリンクした文書の各単語について
7014
COND3の条件満たす？
7015
当該単語および重みをキーワードテーブル19に格納
7016
COND4がルール18に定義されている
7017
対象文書にリンクした文書の各単語について
7018
COND4の条件満たす？
7019
当該単語および重みをキーワードテーブル19に格納
7020

第１１図

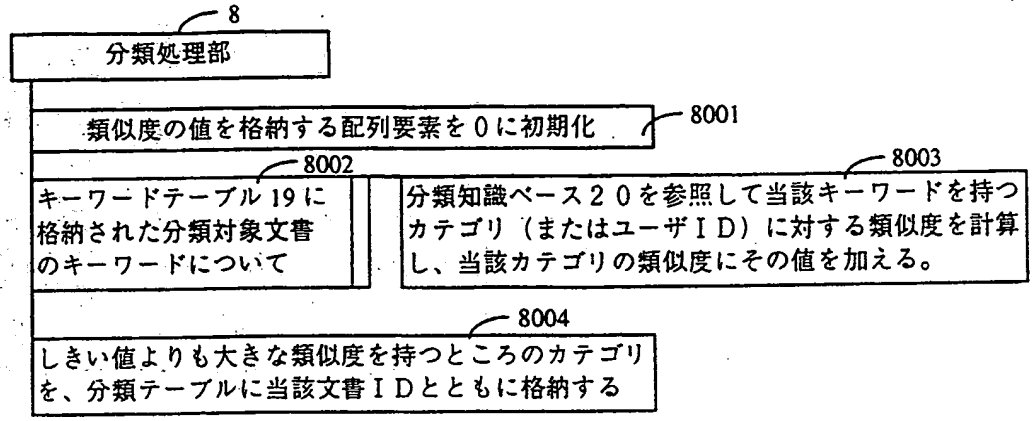
（ａ）カテゴリ分類テーブル

カテゴリ名	キーワード文字列	重み
パソコン	ＰＣ	1.0
パソコン	パソコン	1.0
パソコン	メモリ	0.3
パソコン	・・・・	・・・
ネットワーク	インタネット	1.0
ネットワーク	ネットワーク	0.9
ネットワーク	プロトコル	0.8
ネットワーク	・・・・	・・・
プリンタ	レーザプリンタ	1.0
プリンタ	トナー	0.3
プリンタ	・・・・	・・・

（ｂ）ユーザ分類テーブル

ユーザＩＤ	キーワード文字列	重み
A0001	ＰＣ	1.0
A0001	パソコン	1.0
A0001	メモリ	0.3
A0001	・・・・	・・・
A0002	インタネット	1.0
A0002	ネットワーク	0.9
A0002	プロトコル	0.8
A0002	・・・・	・・・
A0003	レーザプリンタ	1.0
A0003	トナー	0.3
A0003	・・・・	・・・

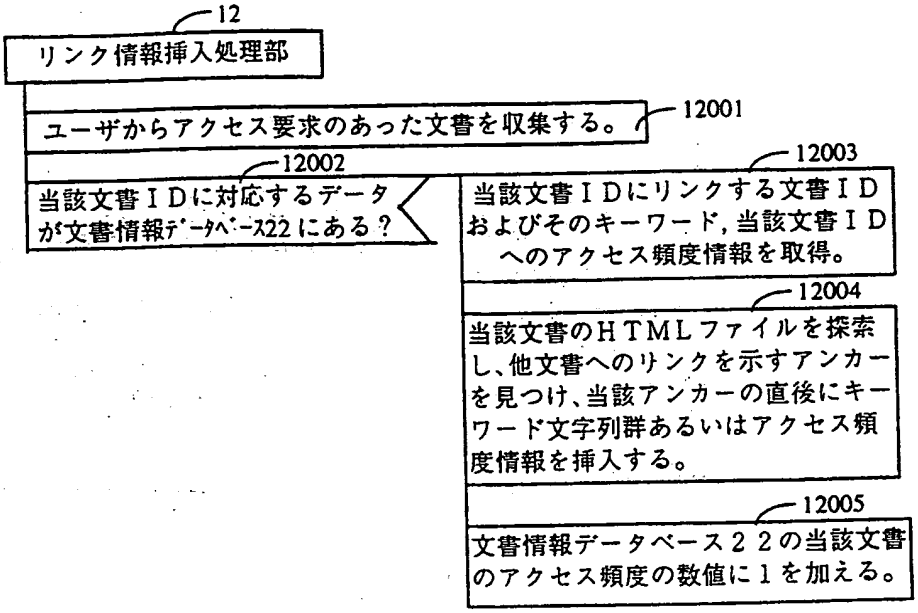
第１２図


8
分類処理部
類似度の値を格納する配列要素を０に初期化
8001
8002
キーワードテーブル19に格納された分類対象文書のキーワードについて
8003
分類知識ベース２０を参照して当該キーワードを持つカテゴリ（またはユーザＩＤ）に対する類似度を計算し、当該カテゴリの類似度にその値を加える。
8004
しきい値よりも大きな類似度を持つところのカテゴリを、分類テーブルに当該文書ＩＤとともに格納する

第１３図

文書ID	更新日	ｶﾃｺﾞﾘ	ｷｰﾜｰﾄﾞ	ｱｸｾｽ頻度	ﾘﾝｸ先文書ID	本文
文書１	1997/1/1	ﾊﾟｿｺﾝ	鶴亀電機(22) ﾎｰﾑﾍﾟｰｼﾞ(22) 会社(22) ・・・	25	文書２ 文書３ 文書４	<HTML> <HEAD> <TITLE> 鶴亀 ・・・・・
文書２	1997/1/1	ﾊﾟｿｺﾝ	ＰＣ(17) 液晶(17) ﾃﾞｨｽﾌﾟﾚｲ(17) ・・・	48	文書４	<HTML> <HEAD> <TITLE> 製品情報 ・・・

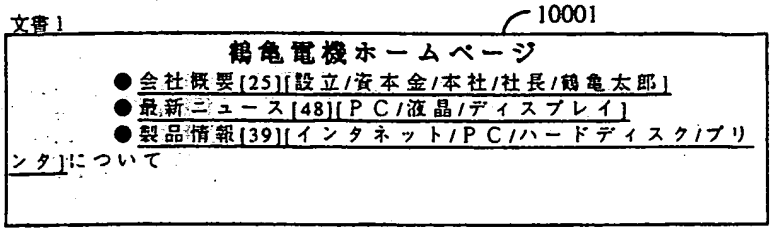
第１４図


12
リンク情報挿入処理部
ユーザからアクセス要求のあった文書を収集する。
12001
12002
当該文書ＩＤに対応するデータが文書情報ﾃﾞｰﾀﾍﾞｰｽ22にある？
12003
当該文書ＩＤにリンクする文書ＩＤおよびそのキーワード，当該文書ＩＤへのアクセス頻度情報を取得。
12004
当該文書のＨＴＭＬファイルを探索し、他文書へのリンクを示すアンカーを見つけ、当該アンカーの直後にキーワード文字列群あるいはアクセス頻度情報を挿入する。
12005
文書情報データベース２２の当該文書のアクセス頻度の数値に１を加える。

第１５図

```
<HTML>
<HEAD>
<TITLE> 鶴亀電機のホームページ</TITLE>
</HEAD>
<BODY>
<H1><B>鶴亀電機ホームページ</B><P></H1>
<A HREF="文書２">●会社概要[25][設立/資本金/本社/
社長/鶴亀太郎]</A><BR>
<A HREF="文書３">●最新ニュース[48][ＰＣ/液晶/
ディスプレイ]</A><BR>
<A HREF="文書４">●製品情報[39][インタネット/ＰＣ/
ハードディスク/プリンタ]</A>について<BR>
</BODY>
</HTML>
```

第１６図


文書1
10001
鶴亀電機ホームページ
●会社概要[25][設立/資本金/本社/社長/鶴亀太郎]
●最新ニュース[48][ＰＣ/液晶/ディスプレイ]
●製品情報[39][インタネット/ＰＣ/ハードディスク/プリンタ]について

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP97/03280

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl^6 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int. Cl^6 G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926 - 1997
Kokai Jitsuyo Shinan Koho 1971 - 1997
Toroku Jitsuyo Shinan Koho 1994 - 1997

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST File on Science and Technology

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	JP, 3-174653, A (Matsushita Electric Industrial Co., Ltd.), July 29, 1991 (29. 07. 91)(Family: none)	1 - 3 4 - 17
Y	JP, 5-20362, A (The Tokyo Electric Power Co., Inc.), January 29, 1993 (29. 01. 93), Par. No. (0017) (Family: none)	4-6, 10-12, 15, 16
Y	JP, 5-342272, A (Fujitsu Ltd.), December 24, 1993 (24. 12. 93)(Family: none)	7 - 13
Y	JP, 64-72231, A (Matsushita Electric Industrial Co., Ltd.), March 17, 1989 (17. 03. 89), Fig. 2 (Family: none)	14 - 16
Y	JP, 8-137893, A (Toshiba Corp., Toshiba Computer Engineering K.K.), May 31, 1996 (31. 05. 96), Fig. 18 (Family: none)	17

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
December 2, 1997 (02. 12. 97)	December 16, 1997 (16. 12. 97)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ９７／０３２８０

A.　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl[6]　Ｇ０６Ｆ１７／３０

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl[6]　Ｇ０６Ｆ１７／３０

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報　１９２６－１９９７年
日本国公開実用新案公報　１９７１－１９９７年
日本国登録実用新案公報　１９９４－１９９７年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

ＪＩＣＳＴ科学技術文献ファイル

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X Y	ＪＰ，３－１７４６５３，Ａ（松下電器産業株式会社）２９．７月．１９９１（２９．０７．９１）（ファミリーなし）	１－３ ４－１７
Y	ＪＰ，５－２０３６２，Ａ（東京電力株式会社）２９．１月．１９９３（２９．０１．９３），【0017】段落，（ファミリーなし）	4-6, 10-12, 15, 16
Y	ＪＰ，５－３４２２７２，Ａ（富士通株式会社）２４．１２月．１９９３（２４．１２．９３）（ファミリーなし）	７－１３
Y	ＪＰ，６４－７２２３１，Ａ（松下電器産業株式会社）１７．３月．１９８９（１７．０３．８９），第２図，（ファミリーなし）	１４－１６

☒ Ｃ欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」先行文献ではあるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０２．１２．９７	16.12.97

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）		
日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００ 東京都千代田区霞が関三丁目４番３号	高瀬　勤　印	5L	9069
	電話番号　０３－３５８１－１１０１　内線　３５６４		

様式ＰＣＴ／ＩＳＡ／２１０（第２ページ）（１９９２年７月）

国際調査報告

国際出願番号 PCT／JP97／03280

C（続き）. 関連すると認められる文献

引用文献のカテゴリー＊	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	JP, 8－137893, A（株式会社東芝、東芝コンピュータエンジニアリング株式会社）31. 5月. 1996（31. 05. 96）, 図18, （ファミリーなし）	17